Barycentric Defense

A V Subramanyam
Department of Electronics and Communication Engineering
Indraprastha Institute of Information Technology
New Delhi, India
subramanyam@iiitd.ac.in

Abhigyan Raj
Department of Computer Science
Indian Institute of Technology
Dhanbad, India
abhigyan.18je0013@cse.iitism.ac.in

***Abstract*—Wasserstein metric based adversarial attacks have attracted a great interest in the recent past. Even though they exhibit strong attacks, surprisingly, they have not been investigated for defense. In this work, we demonstrate that barycenters computed in Wasserstein space can act as a measure of defense against adversarial attacks. We compute the barycenter using marginals obtained from the given image and demonstrate its effectiveness in defense even without any adversarial training. We further analyse the barycenters using GradCam to understand their defensive characteristics. Elaborate experiments on MNIST, Fashion-MNIST, CIFAR-10 and CIFAR-100 demonstrate a significant increase in the robustness of victim classifiers.**

***Index Terms*—Barycenter, Dual Wasserstein, Adversarial defense.**

I. Introduction

DEEP learning systems have shown impressive performance in various applications. However, these systems are vulnerable to adversarial perturbations [1]–[3]. In order to counter these attacks, several defense mechanisms have also been proposed.

In one of the early works, Szegedy *et al.* [4] formulated the adversarial attack as an optimization problem and obtained the adversarial sample using L-BFGS. Several adversarial attacks have been proposed since Szegedy' work [5], [6]. In particular, a vast majority of attacks are proposed using l_p norm only. On the other hand, strong defense measures have been studied in [1], [7]–[9].

Different from l_p norm based attacks, Wasserstein distance based attacks have also been studied in [10], [11]. Unlike pixel-wise distances in l_p space, Wasserstein distance incorporates the geometry of the vectors between which the distance is computed. This makes the Wasserstein space more attractive as minor transformations do not lead to large distances in contrast to other metrics such as Euclidean distance. However, defense based on Wasserstein distance is not yet explored and this is the focus of our work.

In this work, we propose a barycenter based defense in a classification setting. Our hypothesis is that the attacks designed in l_p space may not be effective in the Wasserstein space. In order to verify this, we find the barycenter of the given image using marginals derived from the given image itself. We find that the barycenter is very effective in defending against L_{inf} and L_2 attacks. We also investigate the gradcam [12] images where we demonstrate that the barycenter actually possesses traits of the original image itself which is the reason for effectiveness of barycenter against attacks.

II. Related Works

Adversarial Attacks Some of the robust attacks are iterative FGSM (Kurakin *et al.* [6]), PGD (Madry *et al.* [7]), Carlini and Wagner [13] attacks, Jacobian based attack (Papernot *et al.* [14]), physical attack (Athalye *et al.* [15]), Autoattack (Croce and Hein [2]), momentum based iterative FGSM (Dong *et al.* [16]), single pixel attack (Su *et al.* [17]). These attacks are primarily focused in l_p domain. Attacks in Wasserstein space have also been explored in [10], [11], [18], [19].

Adversarial Defense In response to adversarial attacks, adversarial defense has been proposed to defend the victim models. One of the best defense approach is adversarial training [4], [5], [20]. Madry *et al.* [7] formally studied adversarial training and proposed that such training allows network to defend well against first order adversary. Adversarial logit pairing uses a pair of logits from clean and adversarial examples to defend against adversarial samples [21]. Prominent theoretical studies include TRADES [22] which prove the bounds based on additional regularization term that minimizes the difference in prediction between clean and adversarial samples. In [1], Wong *et al.* proposed to effectively combine FGSM and random initialization to demonstrate better adversarial training. In [23], Zhang *et al.* proposed that adversarial data should not have a uniform importance during training and their effect should be geometrically weighted. GAN based defense approach has been proposed in [24].

Data augmentation has also shown a significant performance improvement against adversarial examples [9], [25]–[27]. Other techniques include enhancement and restoration of images [28], [29], distillation [30], ReLU [31], gradient regularization [32].

Wasserstein Barycenter In the following we discuss Wasserstein distance and barycenter. The entropic regularized Wasserstein distance is given as [33],

$$W_\eta(x,y) = \min_{\Pi} \langle \Pi, C\rangle + \eta\langle \Pi, ln\Pi\rangle$$
$$\text{s.t. } \Pi 1 = x, \Pi^T 1 = y,$$

where, x and y denote probability simplexes, Π denotes the transport plan, C denotes the cost matrix, 1 denotes a vector of ones, η is a regularization parameter, and $\langle ., .\rangle$ denotes inner product.

Let there be N marginals q_k, $k = 1, 2, ..., N$, and barycenter weights β_k. Then the barycenter p is the solution of the following objective [34],

$$\min_p \sum_{k=1}^{N} \beta_k W_\eta(p, q_k).$$

The regularized version of the barycenter is given as,

$$\min_p \sum_{k=1}^{N} \beta_k W_\eta(p, q_k) + J(p)$$

where $J(p)$ can be an l_2 regularizer, for example, $J(p) = \frac{\lambda}{2}\|p\|^2$.

III. Methodology

We discuss our proposed objective formulation in this section. We first formulate a regularized barycenter problem and derive a solution to obtain the barycenter. Let the clean or adversarial sample be x. Let $\mathcal{A}$ be a linear operator, then we define the marginal of x as $q_k = \mathcal{A}_k x$. In our experiments, we use rotation and translation as the linear operator. Then the regularized Wasserstein barycenter is given as,

$$\min_p \sum_{k=1}^{N} \beta_k W_\eta(p, q_k) + J(p), \tag{1}$$

where $J(p) = .5\theta\|x - p\|^2$.

We choose a l_2 regularizer different from the regularizer in [34]. This is because barycenter computation can be unstable when a regularizer is only focusing on p. In our objective function Equation 1, we try that the barycenter is also close to the given sample x in the l_2 space and thus we use $J(p) = .5\theta\|x - p\|^2$. This helps in obtaining a barycenter that can both defend against attacks and does not diverge with increasing iterations.

In order to obtain p, we solve the dual problem of Equation 1 given by,

$$\min_{(u_k)_{k=1}^N, v} \sum_{k=1}^{N} \beta_k H^*_{q_k}(u_k) + J^*(v) \tag{2}$$

where $H^*_{q_k}(u_k) = \gamma[E(q_k) + \langle q_k, \ln K\alpha_k\rangle]$ and $E(q_k) = -\sum_i q^i_k$, $\alpha_k = e^{u_k/\gamma}$.

We can expand $J^*(v)$ as,

$$J^*(v) = \sup_s (v^\top s - J(s))$$
$$= -\|v\|^2/2\gamma + v^\top w.$$

The solution for Equation 2 is given in Algorithm 1.

```
Algorithm 1: Regularized Barycenter
Input: $(u_k, v) \leftarrow 1$ a vector of ones with
       dimension $n \times 1$ , $N$ marginals $q_k$,
       $k = 1, 2, ..., N$, $\beta_k = 1/N$ be the
       barycenter weights, $K = e^{-C/\gamma}$, $\gamma$ =
       0.02, $\theta$ = 1e5, $\tau = 0.2$
Output: $p = \frac{1}{N}\sum_k \nabla H^*_{q_k}(u_k)$
while $max\ iterations$ do
    $u_N \overset{\text{def.}}{=} -\frac{v}{\beta_N} - \sum_{k=1}^{N-1} \frac{\beta_k}{\beta_N} u_k$
    $\nabla H^*_{q_k}(u_k) = \alpha_k \odot K \frac{q_k}{K\alpha_k}$
    $\nabla F((u_k)_{k=1}^N, v) = \big[\big(\beta_k \nabla H^*_{q_k}(u_k) -$
      $\nabla H^*_{q_N}(u_N)\big)_{k=1}^{N-1}, -\nabla H^*_{q_N}(u_N)\big]$
    $x^l \leftarrow ((u_k)_{k=1}^{N-1}, v)$
    $x^{l+1} = Prox_{\tau J^*}(x^l - \tau\nabla F(x^l))$
    $x^{l+1} \leftarrow \frac{x^l - \tau\nabla F(x^l) - \tau w}{1 - \tau/\theta}$
end
```

Here, $Prox_{\tau J^*(v)} \overset{\text{def.}}{=} \arg\min_{v'} .5\|v - v'\|^2 + \tau J^*(v')$.

IV. Experimental Results

A. Implementation Details

The proposed method is implemented in Pytorch framework. In case of MNIST and Fashion-MNIST, we use the architecture proposed in [10]. For CIFAR-10, we use the Resnet-18 pre-trained model available at [35]. MNIST model obtains an accuracy of 99.0% on clean examples, Fashion-MNIST model achieves an accuracy of 90.0%, and, CIFAR-10 model obtains an accuracy of 94.8% on clean examples. For CIFAR-100, we use Resnet-32 with clean sample accuracy of 70.14%. We use the local cost matrix [10] obtained using l_1 norm between (i, j) and (k, l) as $(|i-j|+|k-l|)$ co-ordinates. The size of the local cost matrix is 7x7.

B. Adversarial attack performance

In Table I, we show the performance of clean samples, barycenters and adversarial samples obtained using L_∞ attacks. We compute the barycenters in the following two settings. First, we use 6 marginals comprising of the given image (attacked or clean), $\pm 2^\circ$ rotation of the given image, $\pm 3^\circ$ rotation of the given image and translation by 3 pixels. We refer this setting as $B6$ in Table I. In the second setting, we use the given image (attacked or clean), $\pm 2^\circ$ rotation of the given image, and translation by 3 pixels. We refer this as $B4$. We tried with higher degrees of rotation, however, too much rotation itself degrades the model'performance. Therefore, a rotation of utmost $\pm 3^\circ$ is considered. The choice of translation is also motivated by common data augmentation methods.

TABLE I

Performance of victim model against barycenter computed using l_∞ adversarial samples and clean samples (in %). $\epsilon = 0.17$.

Dataset	L_∞	Barycenter		Clean	Barycenter	
		$B6$	$B4$		$B6$	$B4$
CIFAR-10	0	53.8	55.7	94.8	88.4	90.5
MNIST	0.5	90.3	85.9	99.0	98.9	98.3
F-MNIST	0	67.7	66.9	90.0	89.8	89.7
CIFAR-100	0	21.2	22.3	70.2	60.5	64.6

It can be observed that the accuracy under L_∞ is almost zero. Whereas, in case of barycenters, the accuracy is very high. A similar trend is observed for all datasets as well as for both attacked and clean images. Thus, barycenters can achieve an accuracy close to that of victim model in case of clean images, and also defend the model under adversarial attack.

We show the performance against several attacks in Tables II and III. We use Foolbox [36], [37] to perform the attacks. We report the results in terms of attack accuracy (AA) and barycenter accuracy (BA). AA is the ratio of correctly classified attacked samples to that of the total number of samples. It is computed over the complete test set. We can observe that AA is very low when the adversarial samples are presented to the victim model. On the other hand, when we test using the barycenters generated from these attacked images, the accuracy is high. We denote this accuracy as Barycentric Accuracy (BA).

TABLE II

Performance against attacks on MNIST. $\epsilon = 1$ for L2PGD and L2Carliniwagner. $\epsilon = 0.1$ for Linfpgd, FGSM, Linfdeepfool.

Attack	AA (in %)	BA (in %)
L2PGD [7]	43.59	62.03
L2Carliniwagner [13]	63.19	96.07
Linfpgd [7]	16.36	31.75
FGSM [5]	71.75	74.96
Linfdeepfool [20]	42.52	94.69
AutoAttack [2]	9.13	28.12

TABLE III

Performance against attacks on Fashion-MNIST. $\epsilon = 1$ for L2PGD and L2Carliniwagner. $\epsilon = 0.1$ for Linfpgd, FGSM, Linfdeepfool.

Attack	AA (in %)	BA (in %)
L2PGD [7]	9.45	14.38
L2Carliniwagner [13]	14.14	80.44
Linfpgd [7]	38.39	46.99
FGSM [5]	45.88	51.6
Linfdeepfool [20]	42.59	81.05
AutoAttack [2]	0.6	2.75

We present the results for CIFAR-10 and CIFAR-100 in Tables IV and V, respectively. In both these cases, we find that our model exhibits a strong defense against attacks like L2Carliniwagner [13] and Linfdeepfool [20]. Further, as AutoAttack [2] is quite powerful, it is hard to achieve a sound robustness against this attack without doing any adversarial training. We also performed experiments by increasing ϵ to 0.3 for MNIST and 16/255 for CIFAR-10, however, both attack and barycenter accuracies are low in this case.

Although methods involving adversarial training such as [7], [22] achieve a high accuracy compared to that of ours, the proposed method opens up a new direction for exploring defense without any training.

C. Sensitivity analysis

We show the sensitivity analysis in Figure 1. The image enclosed in red box denotes the barycenter obtained from the settings used for all the experiments. We fix θ and τ, and increase γ to analyse the effect on barycenter. This is shown in the last three columns of first row. We can see that increasing γ deteriorates the quality of the barycenter which can have a negative impact on barycentric accuracy. Similarly, the effect of changing θ

TABLE IV

PERFORMANCE AGAINST ATTACKS ON CIFAR-10. $\epsilon = 0.5$ FOR L2PGD AND L2CARLINIWAGNER. $\epsilon = 8/255$ FOR LINFPGD, FGSM, LINFDEEPFOOL.

Attack	AA (in %)	BA (in %)
L2PGD [7]	2.65	32.51
L2Carliniwagner [13]	0.06	83.07
Linfpgd [7]	0	14.62
FGSM [5]	37.48	43.11
Linfdeepfool [20]	1.81	78.11
AutoAttack [2]	0	10.15

TABLE V

PERFORMANCE AGAINST ATTACKS ON CIFAR-100. $\epsilon = 0.5$ FOR L2CARLINIWAGNER. $\epsilon = 8/255$ FOR LINFDEEPFOOL.

Attack	AA (in %)	BA (in %)
L2Carliniwagner [13]	0.01	55.31
Linfdeepfool [20]	0	54.46
AutoAttack [2]	0	4.54

and τ is given in second and third row respectively. We observe that decreasing θ leads to blurring while increasing θ generates reasonable quality of barycenter. Further, a large τ also deteriorates the quality of barycenter.

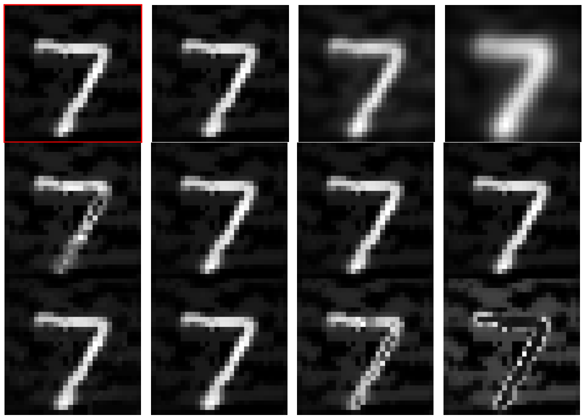

Fig. 1. First row: γ = .02, .01, 1, 5 with $\theta = 1e5$, τ = 0.2. Second row: γ=0.02, $\theta = 1e3, 1e4, 1e6, 1e7$ and τ = 0.2. Last row: γ=.02, $\theta = 1e5$ and τ = 0.005, 0.05, 1, 5.

D. Subjective Evaluation

We show the barycenters in Figure 2. In the figure we can observe that the barycenters obtained under $B6$ setting are close to the respective original images. In Figure 3, we show the gradcam heatmaps. We can observe that the heatmaps for barycenters are closer to the heatmaps of respective original images compared to that of attacked images. Especially, in case of MNIST and Fashion-MNIST, we can clearly observe that heatmaps of the clean images and respective barycenters show a high degree of similarity. In case of CIFAR-10, the heatmap of attacked image is very different from the heatmap of clean image indicating that the model may not be able to correctly classify the attacked image. Thus, empirically we can observe that the barycenter can mitigate the L_∞ attacks effectively without any training.

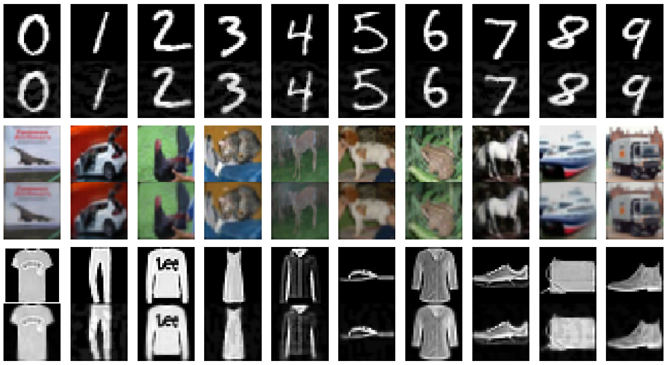

Fig. 2. Clean image, Barycenter ($B6$) for MNIST, CIFAR and Fashion-MNIST.

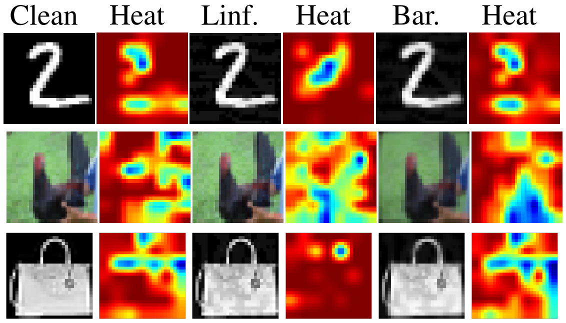

Fig. 3. Gradcam heatmaps of clean images, L_∞ attacked images and Barycenters of MNIST, CIFAR-10 and Fashion-MNIST.

V. CONCLUSION

We present a novel adversarial defense technique and demonstrate that models can be defended without the need of any adversarial training. We formulate a Wasserstein barycenter objective and analytically solve for the barycenter. In order to obtain the barycenter, we use the given image and derive its marginals using rotation and translation. Our empirical analysis using GradCam show that the barycenters possess the traits of clean images and thus have a better probability of being correctly classified. We perform elaborate experiments to qualitatively and quantitatively show that barycenters exhibit a strong defense against a variety of adversarial attacks.

References

[1] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[2] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*. PMLR, 2020, pp. 2206–2216.

[3] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019, pp. 2730–2739.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] R. Theagarajan, M. Chen, B. Bhanu, and J. Zhang, "Shieldnets: Defending against adversarial attacks using probabilistic adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6988–6996.

[9] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Fixing data augmentation to improve adversarial robustness," *arXiv preprint arXiv:2103.01946*, 2021.

[10] E. Wong, F. Schmidt, and Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," in *ICML*. PMLR, 2019, pp. 6808–6817.

[11] J. E. Hu, A. Swaminathan, H. Salman, and G. Yang, "Improved image wasserstein attacks and defenses," in *ICLR 2020*, 2020.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Symposium on security and privacy*, 2017, pp. 39–57.

[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *European symposium on security and privacy*, 2016, pp. 372–387.

[15] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[16] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[17] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[18] K. Wu, A. Wang, and Y. Yu, "Stronger and faster wasserstein adversarial attacks," in *ICML*. PMLR, 2020, pp. 10 377–10 387.

[19] J. Li, J. Cao, S. Zhang, Y. Xu, J. Chen, and M. Tan, "Internal wasserstein distance for adversarial attack and defense," *arXiv preprint arXiv:2103.07598*, 2021.

[20] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, 2016, pp. 2574–2582.

[21] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *NIPS*, 2018.

[22] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*. PMLR, 2019, pp. 7472–7482.

[23] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," *arXiv preprint arXiv:2010.01736*, 2020.

[24] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[25] J. Uesato, J.-B. Alayrac, P.-S. Huang, R. Stanforth, A. Fawzi, and P. Kohli, "Are labels required for improving adversarial robustness?" *NeurIPS*, 2019.

[26] A. Najafi, S.-i. Maeda, M. Koyama, and T. Miyato, "Robustness to adversarial perturbations in learning from incomplete data," *NeurIPS*, 2019.

[27] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi, "Unlabeled data improves adversarial robustness," *NeurIPS*, 2019.

[28] Z. Zhao, H. Wang, H. Sun, J. Yuan, Z. Huang, and Z. He, "Removing adversarial noise via low-rank completion of high-sensitivity points," *IEEE Transactions on Image Processing*, 2021.

[29] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[30] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Symposium on security and privacy*, 2016, pp. 582–597.

[31] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *ICML*. PMLR, 2018, pp. 5286–5295.

[32] A. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *AAAI*, vol. 32, no. 1, 2018.

[33] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *NeurIPS*, vol. 26, pp. 2292–2300, 2013.

[34] M. Cuturi and G. Peyré, "A smoothed dual approach for variational wasserstein problems," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 320–343, 2016.

[35] K. Liu, "Pytorch-cifar," https://github.com/kuangliu/pytorch-cifar.

[36] J. Rauber, W. Brendel, and M. Bethge, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," in *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. [Online]. Available: http://arxiv.org/abs/1707.04131

[37] J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel, "Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax," *Journal of Open Source Software*, vol. 5, no. 53, p. 2607, 2020. [Online]. Available: https://doi.org/10.21105/joss.02607